MEDICAL PRACTICE

Medical Education

Evolution of an Examination: M.R.C.P. (U.K.)

P. R. FLEMING, W. G. MANDERSON, M. B. MATTHEWS, P. H. SANDERSON, J. F. STOKES

British Medical Journal, 1974, 2, 99-107

Summary

Recent changes in the structure of the examination for M.R.C.P. (U.K.) have prompted a review of these together with a historical outline of the examinations which preceded it at the Royal Colleges of Physicians of Edinburgh, Glasgow, and London.

Part I of the examination continues to fulfil its role as a screen and a slight change in the marking system appears to have improved its discriminating powers still further. The major change in part II is the substitution of a more objective written test in place of the essay papers.

Introduction

Interest in the reliability of medical examinations has been smouldering in various corners of Britain since Bull's early article on examinations in 1956.¹ This has been paralleled by the declaration of doubts in other disciplines and the literature is well reviewed by Cox.²

For some reason more attention has been given to reducing the unreliability of traditional examinations in medicine than in other disciplines and many examiners, conscious that their activities would inevitably influence learning processes, had

given much thought to the subject, even before the Todd report³ commented on the need for attention to the quality of examinations. Efforts to improve undergraduate examinations have been visible in patches all over the country, but the most obvious work has been done at postgraduate level, at which objective tests are particularly needed in view of the heterogeneous nature of the groups of candidates, the absence of a comprehensive postgraduate training programme, and the continuing trust placed in British medicine despite the dissolution of the Empire. Backgrounds vary, references are based on variable standards, and more doctors are looking for postgraduate training all over the world. These factors, coupled with the increasing numbers of candidates presenting themselves for the examinations for "membership" have led to substantial changes in examinations which had remained virtually unaltered in London, Edinburgh, and Glasgow for about a century.

Doubtless the pattern will change further in the future and those changes which have already been made will be judged by their impact in the coming years. We have been pressed to put on record some account of the development of the M.R.C.P. examinations up to the present and we feel, now that the new part II M.R.C.P. (U.K.) has been launched, that it is appropriate to review the events which have led up to its introduction, in the hope that it may be of some value to others engaged in a similar exercise.

History of the College Examinations

LONDON

The M.R.C.P. at the London college has always been centred on the clinical examination which is still the one part of the examination at which a pass mark is obligatory. The first examination in April 1859 consisted of two essay papers, a clinical examination, and two oral examinations. Latin, Greek, French, and German translation exercises were included in the written papers. The senior censor's criticism of the variable standard of the examination in 1893 led to the formation of the first com-

Royal College of Physicians of London

P. R. FLEMING, M.D., F.R.C.P.
P. H. SANDERSON, M.B., F.R.C.P.
J. F. STOKES, M.D., F.R.C.P.

Royal College of Physicians of Edinburgh

M. B. MATTHEWS, M.D., F.R.C.P.(ED.), F.R.C.P.

Royal College of Physicians and Surgeons of Glasgow

W. G. MANDERSON, M.D., F.R.C.P.(GLASG.), F.R.C.P.

mittee on the membership examination and a numerical system of marking was introduced. The diploma remained jealously guarded and over the next 15 years some form of examination was demanded even of established figures such as James Mackenzie and Frederick Gowland Hopkins, though it took only the form of an oral comparable with the "final viva."⁴

From 1910 onwards the college was recurrently worried by the thought that those experienced in such newly expanding disciplines as physiology, tropical medicine, and pathology might find it difficult to pass an examination based on proficiency in general medicine and this found its most recent expression in the device in the late 1950s of a part III examination designed to help paediatricians and psychiatrists, from whom less competence in general adult medicine might be acceptable.

Though various alternative paths to the membership, by the submission of published work or by direct recommendation to comitia by the censor's board, came into being the content of the examination did not change greatly. A move to establish it as a threshold examination, designed to evaluate a candidate's fitness for further training rather than a mark of completion of training, disposed of some of the difficulties which specialists had had in the past. Part III was abolished in 1970, by which time the Royal College of Psychiatrists had been founded and was preparing its own examinations, and plans were well advanced for giving special attention to paediatricians in a separate clinical and oral examination in the new Part II in 1972.

Over the years the languages were dropped and in 1963 the second oral or "final viva" was abandoned. It had been recognized for many years that this inquisition conducted by the censors' board did not provide an ideal occasion to make a final judgment on a borderline candidate in view of the tensions it created; few regretted its passing.

In 1944 the examination made its first venture overseas when, as a result of the war and the absence of many young post-graduates on service in the Far and Middle East, a membership examination for candidates in the armed Forces was set up in India and Egypt. This gave the college a taste of the problems involved in setting up an examination overseas and prepared the way for the subsequent administration of the common part I M.R.C.P. (U.K.) in various parts of the world.

In 1946 the numbers of candidates entering for the examination began to increase sharply and, though this was an organizational strain, it may have helped to broaden the examination by resulting in the appointment of a panel of sixteen additional examiners to help the four censors who had until then carried the whole burden of the examination. Now, in 1973, the panel is 105 strong and includes many younger examiners from all parts of the country.

Until 1963 the basic shape of the examination was little different from that of 1859. Clinical and oral examinations were unchanged but the essay papers were less constricting, offering a choice of two from four questions on general medicine and of two from about fourteen questions covering a range of more specialized experience in paediatrics, psychiatry, tropical medicine, dermatology, geriatrics, and other disciplines.

In 1963 the pressure of numbers taking the examination really began to make itself felt. Every candidate had a clinical examination, which he was obliged to pass before proceeding to the next hurdle of the first oral (sometimes known as the "path. viva" as he might be asked to comment on morbid anatomical specimens and sometimes on more outré objects). This put a great strain both on examiners and on those registrars, ward sisters, and secretaries concerned with the hospital arrangements, and indeed on the hospital as a whole, for there was inevitably some interference with its normal work.

It was therefore decided to introduce a screening test as part I of the examination, consisting of multiple-choice questions (M.C.Q.) to be scored by arrangement with the computer laboratory of the University of Newcastle, which already had a suitable program written for undergraduate medical examinations. The censors' board undertook the task of constructing

the questions which formed the first few part I papers. There were only four censors, none with great previous experience of M.C.Q., but control of the M.R.C.P. examination was traditionally and statutorily invested in the censors' board and at first it was considered undesirable to have others contributing to part I.

In 1966, however, the college established the multiple-choice question study group which was charged with the responsibility of submitting part I papers to the censors' board for approval and taking any steps needed to improve the quality of the questions and build up a reliable question bank. This group consisted of 15 men and met eight times a year, once before (a setting meeting) and once after (a review meeting) each examination. The importance of the review meeting soon became apparent. A new computer program allowed for calculation of the discriminatory power of each item* in each question (previously only the correlation coefficient of the question as a whole had been available); this allowed progressive refinement of questions. The group were able to draw on any source they wished for new questions and the whole panel of examiners for the membership was invited to contribute. To familiarize them with the way in which the board and the group approached the problem, guiding bulletins were prepared and meetings of examiners were held regularly in relation to each paper to discuss its content.

The group was anxious to keep part I questions mainly in the clinical area so as to provide something as near as possible to the clinical screen which had been operative in the past. This was not easy to do; new questions from new sources tended to cover the same ground (calcium metabolism and finger clubbing proved to be favourites) and there was a trend, which had to be countered, towards setting questions in peripheral areas of knowledge simply because they had not previously been explored. Nevertheless, the bank grew steadily and in 1967⁵ analysis of the predictive value of part I in terms of the whole examination showed that it was operating satisfactorily.

As a preliminary move towards the new diploma of M.R.C.P. (U.K.), which came into being in 1969, a common part I examination was instituted in 1968. The multiple-choice question group of the London college was dissolved and was replaced by the joint examining board, a body of thirteen men drawn from all three colleges.

EDINBURGH

In 1861 the Edinburgh College of Physicians instituted an order of members, intermediate in rank between licentiates and fellows. Membership was by election until 1880 when the college set up a committee to advise on an examination. Their modified recommendations were accepted in 1881.

The examination consisted of a written paper in the practice of medicine, including therapeutics, and another in "one or more departments of medicine specially professed" as well as practical and oral examinations in each. These selected subjects, as they were to be called, included psychological medicine, general pathology and morbid anatomy, medical jurisprudence, public health, and, surprisingly, midwifery and diseases of women. In 1904 two of particular importance were added, namely, diseases of children and tropical medicine, and later the range was further extended. The council could also exempt from examination candidates aged more than forty and with more than ten years of practice who had proved themselves "highly distinguished for their scientific or practical attainments." In the very early days the number of examinees was small; a man would apply to be examined and the council would

*There is sometimes some confusion in terminology in respect of the parts of multiple-choice questions. The term "stem" is used universally for the initial statement. The separate alternatives are sometimes known, in Britain, as "items" whereas in the U.S.A. the term "item" refers to the whole question and the separate alternatives are known as "completions". In this paper the words "stem" and "item" will be used.

choose a panel of appropriate examiners, a very different state of affairs from the major exercise of later years.

By 1920 the examination lasted for three days. On the first morning, the candidate had a practical examination in the Royal Infirmary for one hour including examination on the use of the sphygmograph, ophthalmoscope, and laryngoscope, and minor examinations of the blood and macroscopic examination of "disease products (sputum etc.)." He was also expected to be able to conduct chemical analysis of urine and to understand "electrical reactions." For the next two hours he would be with a patient and he would write up this "long case" for one and a half hours at the Royal College of Physicians in the afternoon. On the next day he would sit two three-hour papers, one in medicine and therapeutics and another in his selected subject, and on the third day he would have two one-hour oral examinations, one in medicine and therapeutics and the other in his selected subject.

As the number of candidates increased the time devoted to clinical and oral examinations had to be reduced. Between 1935 and 1965 there were increasing numbers of candidates from overseas and in the early 1960s the entry for the Edinburgh examination was comparable with that in London. So great were the numbers that Edinburgh hospitals could not cope with the invasion of candidates and some examiners had to travel to Glasgow and Dundee as well as to non-teaching hospitals in the South-eastern region of Scotland for there to be sufficient patients. The clinical examinations each consisted of a "long case" and "short cases;" the oral examinations were conducted in the hall of the college. There was no restriction, within reason, as to the time devoted to the clinical and oral examinations. Partly in an attempt to relieve pressure on the clinical facilities a part I written screening examination was introduced in January 1966.

It had always been the intention that the selected subject should give a particular opportunity to the physician who had obtained further training in one of the sub-specialties in medicine and, in the early days, there was no reason to doubt that this intention was usually fulfilled. Nevertheless, with time there was a change. Since all candidates had to seek examination in a selected subject, the candidate who had had no opportunity to obtain specialist training had no option but to learn from extra reading in a sub-specialty of his, often rather arbitrary, choice. It was, for example, possible to obtain the M.R.C.P. (Ed.) with cardiology as a selected subject without having seen a cardiac catheterization performed. Such a state of affairs naturally caused concern and, for this and other reasons, the college set up in 1969 an advisory committee on examinations to report on the M.R.C.P. (Ed.) examination and to make recommendations. The deliberations of this committee had only just begun when the recommendations of the London college on changes in their membership examination became available. These recommendations made a strong appeal and the committee recommended that the Edinburgh college should accept them and also continue with negotiations towards a joint examination. This advice the college accepted.

In this way the M.R.C.P. (Ed.) examination came to an end. Naturally many mourned the end of the selected subject and many too felt that the more leisurely pace of the Edinburgh examination allowed for a better assessment of a candidate's ability. On balance, however, the Edinburgh college felt that there was great advantage to be gained from the institution of a common examination, in particular for bringing to an end the undignified march of the less able candidate to as many examinations of the three colleges as he could attend; in addition, a clinical examination in paediatrics, since 1904 a possible selected choice in Edinburgh, was to become a recognized alternative part of the new common M.R.C.P. (U.K.) examination.

GLASGOW

In 1850 the Faculty of Physicians and Surgeons of Glasgow was

allowed, by Act of Parliament, to create Fellows qua physicians or surgeons without examination. Since its foundation, however, the faculty had exercised its right to examine all candidates seeking admission to its membership and decided that it would also award the new diploma of Fellow only after examination. From the earliest days the examination had consisted of written, oral, and practical tests and these continued until 1859 when the examination for Fellowship of the faculty was allowed to lapse because the necessity for examinations was abolished for those on the *Medical Register*.

For the next 27 years the Fellowship was granted to doctors of medicine of universities in the United Kingdom or of foreign universities recognized by the faculty, provided that their candidature was properly proposed and that, at the ballot for their election, they obtained two-thirds of the votes of the Fellows present. Grave dissatisfaction with this method of election led, in 1886, to the institution of a new examination for the Fellowship qua physician. This examination was modelled on that for membership of the Edinburgh College. It consisted of a written paper in systematic medicine including medical pathology and therapeutics, and clinical and oral examinations in medicine. The candidates could "profess" a selected subject in which they sat written, oral, and practical examinations; the selected subjects offered were the same as those offered in Edinburgh. The number of these subjects grew until, in 1924, the list included pathology, bacteriology, midwifery, gynaecology, medical jurisprudence, state medicine, specific fevers, dermatology, psychiatry and neurology, ophthalmology, diseases of children, and venereal diseases.

During the second world war criticism of the proliferation of selected subjects in the Fellowship examination mounted within the faculty. The view was taken that training in a specialty should properly follow training in general medicine and that the option of selected subjects might encourage candidates, without appropriate special training, to regard themselves, nevertheless, as "specialists." In 1946, therefore, the faculty changed the regulations for its examination for Fellowship qua physician so that it consisted of written and oral examinations in the principles of medicine and medical pathology and in the practice of medicine and therapeutics with an examination in clinical medicine.

In 1964 the faculty became, by royal charter, the Royal College of Physicians and Surgeons of Glasgow with the right to grant the diplomas of M.R.C.P., F.R.C.S., and F.R.C.P. No major change in the structure of the examination was needed when the examination for Fellowship qua physician of the faculty became the examination for membership of the college.

During 1965 there were 740 candidates for the Glasgow membership. The examination had to be held in many hospitals in and near Glasgow and the clinical examination of so many candidates placed an intolerable burden on all concerned with its organization. A written examination consisting of twenty questions was introduced in 1966 as a screening test, a pass in this being necessary before the candidate was permitted to proceed to the membership examination proper. This screening test solved the logistical problems of the examination but was regarded as less satisfactory than a wholly objective type of examination. In 1968, therefore, the Glasgow college joined the other colleges in adopting the M.C.Q. examination as part I of the membership examination.

Recent developments in Part I

Until 1969 the multiple-choice questions used in part I were scored and analysed by the methods described by the study group of the London college in 1967.⁵ Early in 1970 a new scoring system was introduced as it was becoming clear that there were certain unsatisfactory features in the original system. One of these was the limitation of the candidate's choice in each item to "true" or "false." It seemed undesirable to force the candidate to guess if the item dealt with an area of knowledge with which he was quite unfamiliar; in an oral examination and

in life in general, a frank admission of ignorance is usually regarded as less reprehensible than a wild, often inaccurate, guess. The answer sheets, which since 1968 had been processed automatically using the English Electric Lector system, were therefore redesigned so that the candidate is now required to make a choice of "true" or "false" for each item by marking the appropriate box on the Lector sheet (fig. 1). By failing to mark either box he can tacitly admit his ignorance of that item.

The introduction of the "don't know" option necessitated a change in the system of scoring. The original system ($+ \frac{1}{p} - \frac{1}{q}$) described by the study group in 1967, though ingenious, was not wholly satisfactory as the candidate's score on each item depended not only on his knowledge but also on the number of positive and negative items in the question. Thus the scoring value of a single item might vary as much as fourfold, introducing a weighting of the items quite unrelated to the intrinsic importance of the knowledge being tested. In the new scoring system one mark ($+ 1$) is given for each item, whether positive or negative, correctly answered, one mark ($- 1$) is deducted for each incorrectly answered item; zero is allotted for each item omitted by the candidate or when, as occasionally happens, a candidate marks an item as both "true" and "false." The candidate's total score on each question is then divided by the number of items in that question, which is now either five or six—unlike the wider range used under the old system. Thus the possible scores on each question range from $- 1$ to $+ 1$, as in the original system.

FIG. 1—Part of Lector sheet.

NEW SYSTEM OF ANALYSIS

Because of the introduction of the "don't know" option and the change in the scoring system, it was necessary to devise a new system of analysis. The discriminatory power of each question as a whole is determined, as before, by calculating the correlation coefficient between all the candidates' scores for that question and their total scores for the whole paper. The devising of a suitable method of analysis of the discriminatory power of the individual items presented some difficulty. In the original system a candidate could be right or wrong only on each item and the analysis of this dichotomy involved the calculation of the point biserial correlation coefficient for each item. Neither this statistic nor the phi coefficient nor any of the other established techniques for the analysis of a dichotomy was immediately applicable to the new system, in which a candidate obtains one of three possible scores on each item. An attempt to use the standard (product-moment) correlation coefficient was soon abandoned when it was recognized as a statistically invalid procedure; it became apparent that the distribution of the item scores was usually far from Gaussian, being often clearly bimodal.

Another approach which was investigated was to omit the candidates who had scored 0 from the analysis of each item and to calculate the point biserial correlation coefficient between

those who scored $+ 1$ and $- 1$; this proved difficult to interpret as the numerical value required to reach significance varied from item to item depending on the number of candidates attempting them. Finally it was decided to calculate the point biserial correlation coefficient using the two groups who were either "right" or "not right" (that is, "wrong" plus "don't know"). This has proved satisfactory in practice and, when compared with results obtained using the original system, has confirmed our suspicions of the shortcomings of the latter; this matter is discussed in more detail below. A portion of the print-out of the item analysis in its current form is shown in fig. 2.

QU.	MEAN	S.D.	NOT	CORR.	ITEM RESPONSE STATISTICS					
1	0.611	0.313	5	0.346	65.375*	78.500	64.063*	84.000*	56.188	% CDS CORRECT
					0.099	0.132	0.247	0.309	0.061	BIS. CORR.
					65.375	1.813	64.063	84.000	7.625	% CDS TRUE
					11.000	78.500	11.688	10.438	56.188	% CDS FALSE
					23.563	19.625	24.125	5.625	36.125	% CDS NOT ATT.
					0.000	0.000	0.000	0.125	0.000	% CDS BOTH MARKED

FIG. 2—Computer print-out of result of analysis of a single question.

COMPARISON OF SYSTEMS

The new scoring system has now been in operation for two and a half years and a comparison with the original system is of some interest. The study group⁵ had demonstrated previously the striking reproducibility of the question mean scores. A similar analysis has now been performed on twenty questions which have been used without modification before and after the introduction of the new scoring system. The two values of the mean scores for these questions gave a correlation coefficient of 0.939, with a regression coefficient of 0.803. A direct comparison of the pairs of mean scores for these questions showed no significant differences for 13 of the questions; for six questions the mean score with the original system was significantly greater than that with the new; the reverse was the case for one question only. The correlation coefficients were nearly all higher when the new system was used but were significantly so for two questions only.

Thus, when the new system of marking was employed, questions tended to have slightly lower mean scores and their discriminatory power was more clearly revealed. The reason for this becomes apparent when individual items rather than whole questions are compared. Out of 123 items which have been used unmodified on at least two occasions before and after the introduction of the new system 63 were positive—that is, "true"—and 60 negative. A comparison of the performances of candidates on these items is shown in the table. For over half the positive items the "% candidates correct" was not significantly different whichever scoring system was used, though in about one-third

Comparison of Percentage of Candidates correct and point biserial Correlation Coefficients for 63 positive Items and 60 negative Items scored and analysed by both the old and the new Systems. The figures refer to the Numbers of Items in each Category

	Percentage of Candidates correct					
	Higher on Old System		No Significant Difference		Higher on New System	
	Pos. Items	Neg. Items	Pos. Items	Neg. Items	Pos. Items	Neg. Items
Point biserial correlation coefficient	1	0	6	0	1	0
	4	38	29	4	20	2
	0	15	1	0	1	1

significantly more candidates correctly marked them as "true" when the new system was used. There seems no very good reason to attribute this to the presence of the "don't know" option and, for some of the items at least, there seems to have

been a genuine increase in knowledge. Nearly all the comparisons of the point biserial correlation coefficient showed no significant difference. A strikingly different pattern was seen for the negative items of which a large majority had a lower "% candidates correct" when the new scoring system was used; for about a quarter the point biserial correlation coefficient was significantly higher with the new system. Probably in the past candidates, by failing to mark an item of which they were uncertain, had in effect correctly indicated it as negative and had obtained credit for knowledge they did not possess. By the introduction of the "don't know" option the true extent of the candidates' knowledge and the discriminatory power of some negative items became apparent.

AMENDED SCORES

No change has been made in the method of producing "amended scores." This process, by which each question is weighted according to its discriminatory power as assessed by the correlation coefficient, was described by the study group.⁵ This procedure was introduced when members of the multiple-choice question group were uncertain of their ability to set highly discriminating questions and its purpose was to ensure that poorly constructed questions would be automatically eliminated from the scoring. This was particularly important as there was, and is, no means of effectively pre-testing the questions before they are used in the examination. Candidates taking part I in the early days sometimes complained informally that some of the questions were ambiguous and it was possible to reassure them that, if the ambiguity became clear on analysis, those questions could have little or no effect on their scores.

With increasing expertise the number of poorly discriminatory questions has fallen to a very low level and possibly the time is approaching when the use of amended scores could be abandoned. The manoeuvre, in any case, has little effect on the result of the examination. The amended scores cover a wider range than the original percentage scores but the means of the two distributions are very similar. Furthermore, a study on a paper used in the L.R.C.P. examination and scored in an identical fashion to the M.R.C.P. has shown no significant difference in the rank order of the candidates whether their scores were amended or not (Spearman coefficient 0.9802; $P < 0.001$). One possible and undesirable effect of using amended scores should be mentioned. Questions in areas outside the mainstream of general medicine, such as dermatology, psychiatry, and tropical medicine, tend to have rather low correlation coefficients and possibly specialists in these fields, who have answered such questions correctly, have failed to obtain due credit for their knowledge.

The possibility of weighting questions according to their "importance" has been considered but a preliminary attempt to produce numerical weights for each question revealed a disappointing lack of unanimity among the members of the joint examining board. The difficulties involved in reaching agreement among specialists on the relative importance of different items of information has been demonstrated clearly by Dudley⁶ and there is no immediate prospect of introducing a comprehensive weighting system into the part I; the possibility of weighting a few questions regarded as of supreme importance is not, however, precluded.

THE ROLE OF THE JOINT EXAMINING BOARD

The main task of the joint examining board is to set examination papers covering the whole field of general medicine and specific attention is paid to ensuring that all specialties are adequately represented in the papers. The board is also responsible for maintaining the standard of the examination at a constant level. Such minor fluctuations as occur in the mean scores of all the candidates from examination to examination are studied to

determine whether they are due to variations in the difficulty of the papers or in the average ability of the candidates. The latter is unlikely to be an important factor as the number of candidates taking the examination is very large (fig. 3) and is, in any case, specifically investigated by the use of "marker" questions. These are questions which have been used in previous papers and have discriminated so well that they can be used again unmodified; it has been shown above that candidates' performances in such questions are nearly always very similar.

FIG. 3—Numbers of candidates entering for part I since 1965.

The Board is also concerned with the predictive value of part I, being particularly anxious, so far as possible, not to exclude candidates who would have been successful in part II if they had been allowed to take it. This subject has been examined on several occasions. In 1967 the study group presented evidence suggesting that, at that time, the part I examination was highly selective. Recently a similar study has confirmed this, though it was not possible to use exactly the same method as in 1967 on account of minor changes in the method of setting the pass mark. A group of 212 candidates who passed part I early in 1971, and who took part II within fifteen months, was divided into subgroups on the basis of their standard scores*. The cumulative percentage pass rates of the various groups at each of the five succeeding part II examinations are shown in fig. 4. The group whose standard score did not exceed the pass mark by more than 0.25 and thus passed part I by a rather narrow margin was clearly less successful in part II than the other groups and it is not unreasonable to extrapolate into the groups who failed part I and to infer that few of them would have been successful in part II.

Most of the time at meetings of the joint examining board is taken up with the construction and amendment of multiple-choice questions. The increasing expertise of the board's members is reflected in the number of questions of which the correlation coefficient is highly significant. The proportion of questions whose correlation coefficient was 0.3 or greater in all the papers since the institution of part I is shown in fig. 5. There was a striking improvement after the appointment of the multiple-choice question group of the London college and a further improvement after this was replaced by the joint examining board. With so many candidates a correlation coefficient as low as 0.08 is usually significant at the 5% level and values above 0.15 reach the 0.1% level.

*The standard score is computed from the expression $\frac{x - \bar{x}}{S}$ where x is the raw value, $\bar{x}$ is the mean of all the raw values, and S is their standard deviation.

FIG. 4—Cumulative percentage pass rate in part II over 15 months of 212 candidates who passed part I in January 1971. These candidates were subdivided according to their marks above pass mark (standard score) as follows: Group 1, 0.0-0.249; 2, 0.25-0.499; 3, 0.5-0.749; 4, 0.75-0.999; 5, 1 and above.

FIG. 5—Proportion of highly discriminating questions in the M.C.Q. paper ($r \geq 0.3$) since the institution of paper that in 1963. M.C.Q. Group = Multiple-choice question study group; J.E.B. = Joint examining board.

A more detailed account of the effects of modification of questions may be of interest. The following question was used in an examination in 1968:

Pseudo-hypertrophic pulmonary osteoarthropathy

- A may respond favourably to vagotomy.
- B may be improved by sympathectomy
- C may be associated with intestinal polyposis.
- D is often associated with high urinary secretion of oestrogens
- E is commoner with peripheral lung tumours than with those of main bronchi.

Analysis using the original system gave the following result (correct items are indicated by an asterisk):

Mean Score	Correlation Coefficient	A*	B	C*	D*	E*
0.262	+0.093	42.2	90.0	26.0	22.8	54.1
	% correct point biserial correlation coefficient	+0.356	+0.009	-0.094	+0.082	-0.067

Overall the question was rather difficult as shown by the low

mean score and discriminated very poorly, the correlation coefficient being barely significant. Inspection of the separate items shows that only A has a significant point biserial correlation coefficient. Several changes were clearly required and when the question was used again in 1973 it was in the following form:

Hypertrophic "pulmonary" osteoarthropathy.

- A may respond favourably to vagotomy
- B is usually painless
- C may be associated with Crohn's disease
- D may be associated with high urinary excretion of oestrogens
- E can be diagnosed radiologically by "tufting" of the terminal phalanges.

The analysis in the new form was as follows:

Mean Score 0-100	Correlation Coefficient 0-364	% correct	A*	B	C*	D*	E
		% correct	40.0	47.7	57.8	20.1	38.0
		% incorrect	36.4	35.9	19.9	25.0	36.0
		% not attempting point biserial correlation coefficient	23.6	16.4	22.2	55.2	26.0
			+0.347	+0.293	+0.135	-0.035	+0.195

The new versions of B, C, and E are clearly a considerable improvement and this is reflected in the highly significant correlation coefficient. The fall in the mean score is largely due to the change in B; possibly among the 90% who answered this item correctly in the original version there were many who would have chosen the "Don't know" option had it been available in 1968. D, in which the changes were trivial, once again failed to discriminate; the fact that 55% of the candidates did not attempt this item makes it clear that this piece of information is far from being common knowledge.

The part I question bank now contains some 3,000 questions; this is due to the continuing efforts of the membership examiners, all of whom are asked to provide a few new questions each time they are asked to examine. The questions are naturally kept under lock and key but the very size of the bank implies that security is not a major problem. Indeed, in 1972 the three royal colleges felt able to authorize the publication of a sample of 200 questions.

The New Part II

In 1968 the London college appointed the latest of a long series of committees to review the membership examination; most of the novel features of the present part II owe their introduction to this committee. One of its main concerns was the problem of examiner variability. This had been well documented in the case of essay questions¹ and some studies on the then existing oral and clinical sections showed that marking standards might vary considerably between examiners.

The organization of the orals and clinicals allowed a fairly simple identification of individual examiners who were marking more harshly or more leniently than the rest. Examiners worked in pairs, each pair interviewing about 35 candidates in the orals and a similar number (of different candidates) in the clinicals. The number of candidates awarded a "fail" mark was noted for each examiner-pair and the result expressed as a percentage of the number of candidates examined by that pair. The results for six examiner-pairs in a particular examination may be quoted as an example:

Examiner-pair	A	B	C	D	E	F
% Failure rate in { Oral	51.3	25.7	36.1	46.7	56.5	26.2
{ Clinical	43.3	24.2	48.8	51.1	43.6	33.3

A high failure rate in the oral tended to be associated with a high failure rate in the clinical, and similarly with low rates. Yet the candidates seen by a given pair of examiners in the oral were not the same as those seen by the same pair in the clinical and no association between the performances of these two groups of candidates would be expected. The correlation coefficient was

calculated for 79 oral/clinical pairings which were studied in this way. The result was $+0.694$, which differs very significantly from the expected value of zero ($P < 0.001$).

HAWKS AND DOVES

By studying the performance of individual examiners in successive examinations (in which they were paired with a series of different colleagues) it was possible to identify a few examiners whose influence appeared to favour high or low marks. For example, one examiner who took part in 10 examinations produced the following results:

	No. of candidates awarded a "pass" mark	No. of candidates awarded a "fail" mark
Pairs containing examiner X	170 (230)	197 (137)
All other pairs in the 10 examinations concerned	1256 (1195)	646 (707)

The figures in parentheses represent the number of candidates expected to receive "pass" and "fail" marks if the distribution of candidates' ability were random, as presumably it was. The χ^2 for this distribution is 51.2 ($P < 0.001$) and there can be little doubt that X was a "hawk" whose influence on his colleagues was such as to lower the pass rate for the candidates he examined substantially below the expected level.

A second study dealt with the concordance or otherwise of marks awarded to the same candidate by two pairs of examiners, together with a "roving" specialist examiner, in the orals. Though the marks in this examination were assigned in the usual way after all five examiners had conferred and agreed a final mark, each examiner was asked to write down his own personal assessment of the candidate before conferring. The pattern of the results obtained is shown in fig. 6; it represents the correlation between the marks awarded to 52 candidates by four "general" examiners, represented by the letters A, B, C, and D. A and B were working together at one table and C and D at another; the candidates spent 15 minutes with one pair and 15 minutes with the other. The only significant correlations are those between examiners working at the same table. The specialist examiner with this group, who was with A and B for part of the examination and with C and D for the rest, awarded marks which correlated well with the mean mark of the pair with whom he was working at the time ($r = +0.778$) and only to an insignificant extent ($r = +0.233$) with the mean mark of the other pair. Moreover, a deliberate attempt was made by the second pair of examiners to avoid topics which had already been discussed by the first pair. Nevertheless, these figures, which are

FIG. 6—Correlation coefficients between marks awarded in an oral examination by four examiners (A, B, C, and D) to 52 candidates.

very similar to those found in two other "quartets" of examiners, indicate that the degree of concordance between two pairs of examiners who interview the candidate separately is poor.

There were thus strong grounds for modifying the examination in such a way as to make the marking more objective and less dependent on the personalities of individuals. In addition clearly the content of the oral and clinical tests varied greatly from one candidate to the next and this constituted a further source of unreliability.

ABANDONING WRITTEN ESSAYS

Quite early in its deliberations the review committee agreed that the written section, with its two three-hour essay papers, was the least satisfactory part of the examination. The main reasons for this conclusion were—firstly, the known unreliability of marking of essay questions, referred to above; secondly, the limited range of subject matter which could be tested; thirdly, the extravagant demands made on examiners' time through the reading of the scripts; and, lastly, the great difficulty in assessing, through the answers, anything other than factual information (which was already being objectively tested over a wide range in part I). The last objection was perhaps the most compelling, arising as it did from actual experience of the examination. Essay questions in medical examinations are defended on the grounds that they provide a means of assessing a candidate's power of original thought, his ability to present a reasoned exposition or argument, and his command of the English language. Nevertheless, these qualities rarely play any part in determining the mark assigned, the literary element in the answers is negligible, and, for the most part, they consist of factual material—sometimes as notes. The Committee realized that the essay questions were being used for a purpose for which they were unsuitable; the introduction of the M.C.Q. test as part I of the examination removed their last justification and it was decided to abandon them.

Much thought was given to the clinical examination, which was clearly unsatisfactory as regards objective marking, and introduced an element of chance in clinical problems which were allotted at random to different candidates. This difficulty seemed inseparable from the traditional British bedside test; as no substitute for this was contemplated, it had to be accepted. Apart from some increased emphasis on testing the candidate's ability to take a good history and to make an assessment of a case as a whole, radical change in the design of the clinical was not thought practicable.

The oral, on the other hand, seemed a hopeful area for improvement. It was clear that the techniques being used and the attributes being tested by different examiners varied widely. Some examiners regarded the test as an opportunity to test the candidate's ability to handle particular clinical situations; others again opened a discussion with the presentation of some laboratory data, an x-ray film, an electrocardiogram, or a clinical photograph. Factual testing in oral examinations is inefficient and unreliable, and in any case extensive factual testing was being carried out in part I. The remaining areas however seemed to be legitimate targets for some rather more objective testing procedure. Only a written test seemed likely to be suitable and, as the essay questions were to be withdrawn, this was clearly the solution.

NEW WRITTEN SECTION

The new written section thus took shape as a deliberate attempt to test what had previously been tested, albeit in a haphazard fashion, in the oral—problem solving, interpretation of laboratory data, and visual recognition. A wide range of test material could be employed in a single examination, with a resulting increase in reliability. As regards scoring, the most convenient and most objective technique would have been a multiple-

choice format, but this was rejected because of the inevitable "cueing" effect of such questions. It was felt that in—for example, a diagnostic problem, it was important to find out whether the candidate could think of a diagnosis or an appropriate investigation by himself, without being prompted from a list of alternatives. The solution finally adopted was hand-marking of answers restricted to a few words only, the acceptable responses having been agreed previously and marks being awarded according to a strict scheme of ratings. It seemed that there might be a case for dropping the oral altogether, but the importance of personal confrontation with the candidate was also recognized and, for this and other reasons, a modified oral was proposed.

These ideas were well received by the London college, and it was agreed to carry out a practical trial of the new written section, which involved quite unfamiliar techniques and which might well contain hidden pitfalls. For this purpose 17 medical registrars kindly agreed to act as "candidates," and this trial examination was held in the London college in July 1969. It was apparent that the tests proposed were capable of realization. Those concerned with planning the new examination were heartened by the response of the "candidates," who already had had experience of the old-style examination. Asked whether they would have preferred to be examined by the new tests rather than the old, all but one said that they would have preferred the new tests. Arrangements to introduce the new examination were put in hand.

Up to this point the concept had been simply of an improvement in the London examination. Nevertheless, the new techniques proposed for the London part II aroused much interest in the Edinburgh and Glasgow colleges and, by agreement between the three colleges a part II planning committee was set up, charged with the task of preparing and organizing the first four complete written part II examinations. This committee subsequently became the common examining board and was given the responsibility for continued scrutiny of the whole of the part II examination.

UNIFORM STRUCTURE

An important achievement was uniformity of structure of the part II examination at all three colleges. This involved some far-reaching concessions by individual colleges, the most important of which was the abandoning by the Edinburgh college of the "selected subject." The marking structure of the London examination was also modified so as to bring it into line with those in use at Edinburgh and Glasgow.

The committee also worked out the details of collecting material for examination questions, reviewing this critically in small committees based on the individual colleges, and assembling it into examinations which seemed to be of the right level of difficulty and to have a suitable balance of subject-matter. Arrangements were made with the National Foundation for Educational Research for item analysis of each examination by computer. The pattern of setting and review meetings, already established for part I, was adopted for part II also.

The trial examination with 17 registrars had shown that the techniques of the written section were practicable but gave no indication of how they would work with larger numbers. A further trial was therefore staged; this time Fellows of all three colleges consented to serve as "candidates" and 103 took part. The results confirmed that the new test was a viable proposition and gave valuable insight into problems of printing, presenting, and scoring the examination material.

THE PRESENT WRITTEN SECTION

The present written section falls into three subsections, case histories, data interpretation, and projected material. In the first of these candidates are presented with a set of five case

histories, of which at least one deals with a paediatric problem; they must answer questions about three of these. To ease the labour of hand marking, the questions are phrased so as to elicit very brief answers—for example, "list three diagnoses which you regard as probable," "suggest four additional laboratory investigations which you think would be helpful in diagnosis," and so on. The marking structure can be made to reflect the examiners' views as to the importance of the possible answers—for instance there might be only one really acceptable diagnosis and this would probably have a mark of 5 assigned to it. Others less likely might be rated 3, 2 or 1; irrelevant suggestions would score 0. The marking is so arranged that the maximum possible score is the same regardless of which three histories are selected.

The data interpretation section consists of 10 questions, all of which must be answered. The material used includes biochemical and haematological tests, electrocardiograms, lung function studies, and so on; the questions again call for brief answers.

In the projected material section slides of clinical photographs and of x-ray films are projected in a lecture theatre. There are 20 slides and about a quarter of these are of x-ray films. Each slide is shown for two minutes and candidates are asked to write down answers on the combined question and answer sheet provided. Careful attention has been paid to the problem of ensuring that viewing conditions are satisfactory for all candidates at any one centre and are of identical standard at all three centres.

THE PRESENT ORAL EXAMINATION

Much of the ground covered by the old oral test is now tested in the new written section. The oral test has been retained because the personal confrontation involved is thought to be valuable and because of a need to test a candidate on the management of medical emergencies, the basic scientific principles on which modern medicine is based, and the planning of schemes of therapeutics and management to deal with specific clinical situations, including their sociological and psychological aspects.

THE PRESENT CLINICAL EXAMINATION

The clinical test in the M.R.C.P. has, for many years, conformed to the familiar pattern of a "long case" and several "short cases." In the past the range of patients presented as "long cases" has been somewhat restricted, being confined to those with long-standing chronic conditions such as rheumatic heart disease, rheumatoid arthritis, multiple sclerosis, cirrhosis of the liver, chronic leukaemia, and the like. This limitation no doubt arose from the bias towards patients with abnormal physical signs, so that they could be presented either as "long cases" or as "short cases." The presentation of such patients as "long cases" is unsatisfactory because of the tendency to call upon the same patients repeatedly so that the history becomes stereotyped and the exercise becomes a poor test of a candidate's skill. A much wider variety of patients can be used if the need for physical signs is set aside; important historical material with few or no abnormal signs may be presented by patients suffering from such conditions as angina pectoris, peptic ulcer, ulcerative colitis, depression, and so on. Accordingly, a candidate may now find that all the important information about his "long case" arises from the history and that abnormal physical signs contribute nothing to the diagnosis, apart from the fact that they are absent. Candidates are warned of this possibility and are told that if there are no abnormal signs they will be given credit for saying so. In this way the "long case" can be made a more realistic test of the sort of clinical encounter which actually occurs in practice.

Testing the ability to elicit, recognize, and interpret abnormal signs is carried out with the "short cases" and a deliberate effort is made to test over a wide range.

The Future

The functioning of the common part I M.R.C.P. (U.K.) is of continuing concern to the joint examining board. Though it retains a clinical flavour, it is generally conceded that what it does is to test a candidate's fund of information in the cognitive domain, usually at a low taxonomic level. Such knowledge is essential for medical practice but may be less important than the ability to solve problems and to take decisions. A clinically inadequate candidate can familiarize himself by a sheer effort of memory with enough factual knowledge to pass part I yet fail badly in part II. This results in improper use of examiners' time in evaluating candidates whose standard of clinical performance gives them little chance of a pass mark and, even more important, introduces a danger of squeezing out of the list of those who pass part I some good clinicians as a result of crowding of the upper end of the rank order by those who have little more to offer than a good memory. This issue is kept under constant review by the joint examining board.

The question arises whether the new written part II, if it proves its worth, could provide a more reliable screen for the future. It will take some time before its merits can be judged and doubtless the quality of test material will improve over the years as occurred in part I. Even if it is accepted, however, that the content of the written part II in time will reach a level of reliability which inspires confidence, there remains the great difficulty of the logistics of scoring. At present, scoring is carried out by groups of examiners sitting together with the opportunity to discuss and agree the weighting to be given to those answers which have not been foreseen by the common examining board. This works quite well for the numbers now examined in part II at the three colleges, but would put a distinct strain on examiners if the written part II were used as a screen for the larger numbers of candidates who now enter for part I. Nor would it be easy to ensure the comparability of the projected section if it were to be mounted in the large number of centres in Britain and overseas at which part I is now taken. Projection difficulties could be overcome if colour photographs and x-ray photography were reproduced in a booklet but, quite apart from increased costs, many members of the board are, as yet, not satisfied that printed material of this kind is of sufficient quality to provide as fair a test as projection on to a screen.

If it became feasible to produce a written part II in booklet form in the future, the problem of hand-scoring would remain unless the questions asked were framed in M.C.Q. form which could be computer scored. There is no technical difficulty in doing this, but it has to be accepted that there are quite serious objections to "cueing" answers in a clinical problem-solving context.

The present oral provides an opportunity to keep in front of the candidate the importance of taking a holistic view of medicine, at least in the early part of his postgraduate training, since he may be faced with, for instance, a psychiatrist, a dermatologist, or a tropical medicine expert; it also gives an opportunity to examiners to seek and to reward, when found, any indication of imagination or initiative. Until more objective ways are found to assess the attitudes and personal qualities which are so important in most branches of medical practice, the oral may be defended as giving another opportunity, supplementing the clinical confrontation, to make a judgement in this area.

The decision to use the "long case" in the clinical examination

for testing specifically the crucial art of history-taking has produced problems for examiners and for registrars organizing examinations. Examiners are already getting used to their new role, though the structure of the examination does not allow as much time as could be wished for listening to the candidate at work in the way that he is commonly watched at work when eliciting physical signs on an extended range of short cases. Registrars' problems arise from the more patients needed in each examination, which implies a need for more secretarial help in the preparation of case notes and for more space in which to accommodate patients. Few hospitals have a separate examining suite within their curtilage and interference with teaching and with the admission of patients is bound to occur. This can be answered up to a point by holding the examination at more hospitals, not necessarily university hospitals.

The possibility of providing clinical examiners with a check list of itemized areas in which they should test will doubtless be discussed in the future—for example, in a "long case," "has the candidate taken an adequate family history or social history?"; "how do you rate his general approach to the patient?"; and, on a cardiac "short case," "can he locate the apex beat?," "is he able to report on the jugular venous pulse?"; the answers to these questions could be expressed on a numerical scale. For the next few years, however, it will be best to allow examiners to become comfortable with the changes which have already taken place.

The problem of variability of marking standards between examiners remains. A start has been made in identifying "hawks" and "doves" and examiners will be given an opportunity of discovering, in confidence, their avian characteristics.

It is difficult to ignore the growing use of the common part I by overseas countries (quite apart from the Irish college which uses this screen test, on an agency basis, accepting the British passmark for entry into the M.R.C.P. (I) part II). The examination can now be taken, once a year, in Egypt, Ghana, Iran, Malaysia, Sri Lanka, and the West Indies and applications are being considered from other countries. It is convenient for these countries to be able to judge, to some degree, the quality of their postgraduates at home without having to go to the expense and trouble of sending them to Britain. Undoubtedly, however, they would do better, in time, to organize their own postgraduate examinations, giving special weight to medical problems of particular local importance which might not receive emphasis in Britain. It is partly in the hope that it may help those who undertake such a task in the future that this account of the evolution of the M.R.C.P. (U.K.) has been written.

Requests for reprints to P.R.F., Central M.R.C.P. (U.K.) Office, c/o Royal College of Physicians, 11 St. Andrew's Place, London, NW1 4LE.

References

- ¹ Bull, G. M., *Lancet*, 1956, 2, 368.
- ² Cox, R., *Universities Quarterly*, 1967, 21, 292.
- ³ Royal Commission on Medical Education, 1965-8. *Report*, Cmnd. 3569. London, H.M.S.O., 1968.
- ⁴ Cooke, A. M., *History of The Royal College of Physicians of London*, London, Oxford University Press, 1972.
- ⁵ Study Group of the Royal College of Physicians of London, *Lancet*, 1967, 2, 1034.
- ⁶ Dudley, H. A. F., *British Journal of Medical Education*, 1969, 3, 155.